DOCUMENT RESUME

ED 357 059 TM 019 785

AUTHOR Hambleton, Ronald K.; Jones, Russell W.
TITLE Item Parameter Estimation Errors and Their Influence on Test Information Functions.
SPONS AGENCY Graduate Management Admission Council, Princeton, NJ.
PUB DATE Apr 93
NOTE 31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Simulation; *Error of Measurement; *Estimation (Mathematics); Item Banks; *Item Response Theory; *Sample Size; *Test Construction; Testing Problems; Test Items
IDENTIFIERS Accuracy; Graduate Management Admission Test; *Information Function (Tests); Item Parameters

ABSTRACT

Errors in item parameter estimates have a negative impact on the accuracy of item and test information functions. The estimation errors may be random, but because items with higher levels of discriminating power are more likely to be selected for a test, and these items are most apt to contain positive errors, the result is that item information functions and corresponding test information functions tend to be inflated in relation to their true values. The impact of this "capitalization on chance" in item selection on the accuracy of test information functions was investigated, using data from the 1985 administration of the Graduate Management Admissions Test with 2 sample sizes (500 and 2,000). Two factors seemed especially important in determining the size of impact: examinee sample size used in calibrating test items, and the ratio of item bank size to the length of the test constructed using items from the bank. Results from computer simulation clearly indicate that both factors influence test information accuracy, often substantially, with serious problems in accuracy arising when test items were calibrated on modest sample sizes (n=500) and when test item banks were large in relation to the number of items in the tests being constructed. Four figures illustrate the discussion, and three tables present analysis results. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED357059

Item Parameter Estimation Errors and Their Influence on Test Information Functions

Ronald K. Hambleton and Russell W. Jones
University of Massachusetts at Amherst

Abstract

For test developers working within an item response theory framework, the concepts of item and test information are important and useful. Unfortunately, errors in item parameter estimates have a negative impact on the accuracy of item and test information functions. The estimation errors may be random, but, because items with the higher levels of discriminating power are more likely to be selected for a test, and these items are most apt to contain positive errors, the result is that item information functions and corresponding test information functions tend to be inflated in relation to their true values. The purpose of this paper was to investigate the impact of this "capitalization on chance" in item selection on the accuracy of test information functions. Two factors seemed especially important in determining the size of the impact: (1) examinee sample size used in calibrating test items, and (2) the ratio of item bank size to the length of the test constructed using items from the bank. The results of the study were clear: both factors influenced test information accuracy, often substantially, with serious problems in accuracy arising when test items were calibrated on modest examinee sample sizes (N=500) and test item banks were large in relation to the number of items in the tests being constructed.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RONALD K. HAMBLETON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM019785

BEST COPY AVAILABLE

Item Parameter Estimation Errors and Their Influence on Test Information Functions[1,2,3]

Ronald K. Hambleton and Russell W. Jones
University of Massachusetts at Amherst

The number of test developers using item response theory (IRT) models and methods in their test development and related technical work has increased substantially in the last 15 years (Hambleton, 1989; Hambleton & Swaminathan, 1985; Lord, 1980). Item response theory, particularly as reflected in the one-, two-, and three-parameter logistic models for dichotomously scored items, is receiving increasing attention from test developers in test design and test item selection, in addressing the detection of biased test items, in computer-administered testing, and in the equating and reporting of test scores. Nearly all major test publishers, state departments of education, and large school districts currently use IRT models in some capacity in their testing work.

A problem that arises when applying IRT models in test development involves "capitalizing on chance" due to positive errors in some item parameter estimates. The problem arises because test developers, not surprisingly, prefer to select test items, other factors aside, with the highest discrimination indices. But high discrimination indices, on the average, are spuriously high because of positive errors in the item

---

[1]The research described in this paper was funded by the Graduate Management Admission Council. The GMAC encourages researchers to formulate and freely express their own opinions. The opinions here are not necessarily those of the GMAC.

[2]Laboratory of Psychometric and Evaluative Research Report No. 212. Amherst, MA: University of Massachusetts, School of Education.

[3]Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, Georgia, April 1993.

parameter estimates. As a result, the measuring power of tests is overestimated and the errors associated with ability estimates are correspondingly underestimated (if the inflated item parameter estimates are used). The practical consequence is overconfidence in the ability estimates when confidence bands are set (Tsutakawa & Johnson, 1990). The same problem arises in computer-adaptive testing which is one of the most popular and important uses of IRT. The problem may also help to explain the shortcomings of pre-equated tests. Quite simply, the original item discrimination parameter estimates may be overestimates of the true values and tend to be lower, on the average, when they are recalibrated at a later time.

"Capitalizing on chance" was well known in classical measurement (Gulliksen, 1950) within the context of using point biserial and biserial correlations in item selection. The most discriminating items in a field-test administration tended to discriminate less well in the actual test administration. Of course, this is because the first estimates tended to be over-estimated due to positive errors in the estimates. The problem is also well-known in the context of regression analysis. Here, it is common to assess the merits of a regression equation in a cross-validation sample (to minimize the problem of "capitalizing on chance"), and where formulas to predict shrinkage in multiple correlations due to "capitalization on chance" abound. But, to our knowledge, the problem has not been discussed with IRT item parameter estimates except in one earlier paper of ours (Hambleton, Jones, & Rogers, in press).

The purpose of this paper was to investigate the impact of "capitalizing on chance," which arises in item selection, on the accuracy of test information functions. Two factors seemed especially important in

determining the size of the impact: (1) examinee sample size used in calibrating test items, and (2) the ratio of the size of the item bank to the length of the test constructed using items from the bank. Therefore, these two factors were investigated in the study.

Computer simulation methods were used in the research and the characteristics of the item bank were based on item parameter estimates obtained from administrations of the Graduate Management Admissions Test. The first factor seemed important because sample size is inversely related to item parameter estimation errors. With modest-sized samples, item parameter estimation errors are larger, and the possibility is greater for capitalizing on chance in item selection. The second factor was included because the ratio of item bank size to test length also seemed like it would affect the accuracy of test information functions. In general, the larger the item bank and the shorter the test of interest, the more opportunity to "capitalize on chance" by selecting spuriously high discriminating items. The consequence is that, again, test information functions would be misleading.

Since it is rarely the case that only statistical criteria are used in item selection, the study was carried out twice: first, using only statistical criteria in item selection (i.e., optimal item selection), and second, using both statistical and content criteria (i.e., content optimal item selection).

Method

This simulation study was based on item statistics obtained from the 80 item problem solving subtest of a 1985 administration of the Graduate Management Admissions Test (GMAT) (Kingston, Leary, & Wightman, 1988). The

three parameter model was used in data simulation because this model fits many datasets (see, for example, Kingston, Leary, & Wightman, 1988) and because item parameter estimation errors were expected to be larger than item parameter estimation errors with simpler models. Our interest was in assessing the magnitude of the inflation of test information functions in some common situations, hence the reason for our interest in the three-parameter model.

Variables in the Study

This research investigated the influence of two factors: examinee sample size used in item parameter estimation and ratio of item bank size to test length.

Sample Size. Two sample sizes were chosen, N = 500 and N = 2000. Research has shown that samples as small as 500 are just below the minimum sample size recommended for use with the three-parameter model (Hulin, Lissak, & Drasgow, 1982). A sample size of 2000 is generally considered to be larger than is needed in practice to obtain satisfactory item parameter estimates.

Ratio of Item Bank Size to Desired Test Length. The item bank contained 80 items to match the GMAT problem solving subtest. With an upper limit of 80 items there existed practical limits on the ratios which could be effectively investigated. There is some evidence in the literature that ratios as high as 12:1 may be used in practice. Indeed, Bezruczko and Reynolds (1987) report a ratio of 16:1. For the purposes of this study we selected the mid range ratios of 8:1 and 4:1. The influence of ratio of item bank size to test length was investigated by creating either a 10- or 20-item test from the 80 available items. In this way ratios of 8:1 (80:10) and 4:1 (80:20) were obtained.

Item Selection Method

Two methods of item selection were considered in the study. Both are IRT-based; the first is focused on item statistics only; the second is focused on both item statistics and item content.

Optimal Item Selection. Optimal item selection is a method of selecting items from an item bank based on the principles of IRT. Items are selected based on their capability of providing maximum information at a specified point (or range) along the ability scale which is of interest to the test developer. Recently, computer software has become available (Verschoor & Theunissen, 1991) that permits item selection to be done via optimizing algorithms (Theunissen, 1985, 1986; van der Linden & Boekkooi-Timminga, 1989). Desirable test characteristics can be specified (such as test length and the target information function) and then the algorithms are initiated to select the set of items to meet the specified test characteristics. This software, entitled Optimal Test Design (OTD), was used to select test items to meet several of the current GMAT statistical specifications.

Content Optimal Item Selection. Most tests are constructed with both statistical and content specifications. In this simulation, items were sequentially assigned to one of four content categories (1, 2, 3, or 4). To enable the impact of content specifications to be studied, it was necessary to insure that item statistics varied somewhat across content categories. This was accomplished by reassigning several highly discriminating items to alternate content categories to insure item statistics were not identical in the content categories. Means and standard deviations for the item parameter estimates within each content category were:

Examinee Sample Size	Category	Item Parameter b $\bar{X}$	b SD	a $\bar{X}$	a SD	c $\bar{X}$	c SD
2000	1	.16	1.14	.85	.35	.22	.07
	2	-.60	1.36	.62	.20	.22	.08
	3	.30	.99	.89	.26	.19	.08
	4	.56	.81	.86	.23	.18	.07
500	1	-.07	1.02	.85	.28	.20	.06
	2	-.22	1.17	.68	.22	.22	.06
	3	.33	1.14	.89	.26	.20	.07
	4	.67	.87	.84	.22	.18	.05

The influence of content optimal item selection was investigated by requiring OTD to select items under the condition that 40 percent of items were drawn from category 2 and 20 percent from each of categories 1, 3, and 4. This particular allocation had no special significance. Our intent was simply to force content-like specifications (or restrictions) on the item selection process.

Procedure

The specific steps of the simulation were as follows:

1. An examinee item calibration sample was chosen (N = 500 or N = 2000).

2. The computer program DATAGEN (Hambleton & Rovinelli, 1973) was used to simulate ability scores (normally distributed with mean = 0 and standard deviation = 1) and item responses for examinees on all 80 items (referred to as Sample A). True item parameters were taken from the report by Kingston, Leary, and Wightman (1988).

3. BILOG (Mislevy & Bock, 1991) was used to obtain Sample A three-parameter model item parameter estimates for the 80 test items.

4. Steps 2 and 3 were repeated for a second sample of randomly equivalent examinees (referred to as Sample B) to obtain Sample B item parameter estimates.

5. A test length was chosen (either 10 or 20) to vary the bank size to test length ratio

6. An item selection procedure was chosen (either optimal or content optimal).

   Optimal. The test information function for the 80 item GMAT subtest was used as a basis to create target information functions. When a 20 item test was constructed the specified target information function was one quarter that of the GMAT test information function. When a 10 item test was constructed the specified target information function was one eighth that of the GMAT test information function. OTD was used to select items.

   Content Optimal. The same procedures were used to specify target information functions as those used in the optimal item selection procedure. OTD was used to select items specifying the selection of 40% of items from content category 2 and 20% from each of content categories 1, 3, and 4.

7. A test was constructed from the item bank containing the Sample A item statistics. Then the test information functions were calculated using the Sample A item statistics and the corresponding (cross-validation) Sample B item statistics. The relative efficiency function using the Sample B-based test information function as the baseline was also calculated.

8. We repeated the previous steps for (a) item banks calibrated with two sample sizes, (b) two test lengths, and (c) two item selection methods.

The result of steps 2 to 4 was the production of four banks of 80 items. All items in the banks were described by three-parameter model item statistics. In two banks the item statistics were based on relatively small samples (N=500), and in the other two banks the item statistics were based on relatively large samples (N=2000). The next phase of the work involved test development (or, more specifically, item selection). All of the item selection work was done using Sample A item statistics and the test information functions were compiled using Sample A item statistics. For cross-validation purposes, test information functions were also compiled using the corresponding Sample B item statistics.

Results

Descriptive statistics of the item parameter estimates in the four 80-item banks (N=500 or 2000, Sample A or B) are reported in Table 1.

Insert Table 1 about here

Using common item equating, the Sample B item parameter estimates were slightly adjusted to be on the same scale as the corresponding Sample A item parameter estimates (Hambleton & Swaminathan, 1985). As is clear from Table 1, only slight adjustments to the Sample B item parameter estimates were necessary. These adjustments were made prior to completing the main part of the study.

Optimal Item Selection

Descriptive statistics for the 10- and 20-item tests constructed using optimal item selection with the Sample A item banks calibrated with 500 and 2000 examinees, respectively, are reported in Table 2. The corresponding descriptive statistics substituting the Sample B item parameter estimates are also given in Table 2. Figure 1 shows the test information functions

---

Insert Table 2 about here

---

for the various constructed tests. Relative efficiency functions comparing Sample A with Sample B results are shown in Figure 2.

The most important observation in Table 2 is that "capitalization on chance" is operating with the item discrimination values. Items selected for tests using the Sample A item parameter estimates because of their high discriminating power do not function nearly as well in the cross-validation

samples. The average differences ranged from .05 to .11 across the four tests that were constructed. More detailed analyses follow next.

Insert Figures 1 and 2 about here

Sample Size. Figure 1 shows examinee sample size to be an influential factor in the overestimation of test information functions. Clearly the Sample A test information functions were overestimated, and the magnitude of this overestimation was much greater in the smaller calibration sample size (N = 500) than the larger calibration sample size (N = 2000) regardless of test length. Because of the extreme instability of the test information and relative efficiency functions when low amounts of information are present, only comparisons when test information exceeded 2.0 were considered. The relative efficiency functions were calculated over the range for which test information exceeded a value of 2.0.

The two main findings with respect to sample size are:

1. With a ratio of bank size to test length of 8:1 and a modest item calibration sample size (N = 500), the test information function was overestimated by as much as 40 percent. With an identical ratio of bank size to test length and a larger item calibration sample size (N = 2000) the overestimation, although still present, did not exceed 25 percent, and was generally lower over the region on which relative efficiency was calculated.

2. With the smaller ratio of bank size to test length of 4:1, overestimation of the test information function occurred but the magnitude of the overestimation was reduced, as it should have been since there was less opportunity to capitalize on only the outliers. With the modest item calibration sample size (N = 500), the overestimation did not exceed 36 percent, whereas with the larger calibration sample size (N = 2000) the overestimation did not exceed 13 percent, and again, was generally lower over the region on which relative efficiency was calculated.

Ratio of Item Bank Size to Test Length. Figure 1 shows the ratio of item bank size to test length to be another influential factor affecting the magnitude of overestimation of test information functions. Though for the conditions simulated, the impact was less than sample size. The two main findings were:

1. With the modest sample size (N = 500) and a bank size to test length ratio of 4:1, the magnitude of overestimation was as great as 36 percent. However, when the ratio was 8:1 the magnitude of overestimation was as great as 40 percent.

2. Similar trends were present with the larger calibration sample (N = 2000). Again, the magnitude of the difference was larger when the ratio was 8:1 compared to 4:1. However, overall the differences were smaller than those found from the more modest item calibration sample size.

The relative efficiency functions shown in Figure 2 provide the results graphically. Overestimation was greatest when smaller examinee samples were used in item calibration; overestimation was least when larger examinee samples were used in item calibration and the ratio of bank size to test length was low (i.e., 4:1).

Content Optimal Item Selection

Descriptive statistics for the 10- and 20-item tests constructed using content optimal item selection with the Sample A item banks calibrated with 500 and 2000 examinees, respectively, are reported in Table 3. The corresponding descriptive statistics substituting the Sample B item parameter estimates are also given in Table 3. Figure 3 shows the test

---

Insert Table 3 about here

---

information functions for the various constructed tests. Relative efficiency functions comparing Sample A with Sample B results are shown in Figure 4. A similar pattern of overestimation to that exhibited with the

---

Insert Figures 3 and 4 about here

---

optimal item selection procedure is apparent with the content optimal item selection procedure. However, comparisons between Figures 1 and 3 and Figures 2 and 4 show the magnitude of overestimation to be substantially greater in tests constructed using content optimal item selection procedures. No generalizations of this finding should be made, however, since the finding is likely specific to the item banks and content specifications that were used in the study.

Table 3 highlights some substantial effects due to "capitalizing on chance." These effects were especially significant in the item bank calibrated with the smaller examinee sample size.

Sample Size. Clearly Sample A test information functions in Figure 3 were overestimated, and the magnitude of this overestimation was much greater in the smaller calibration sample size (N = 500) than the larger calibration sample size (N = 2000). The two main findings with respect to sample size were:

1. With a ratio of bank size to test length of 8:1 and a low item calibration sample size (N = 500), the test information function was overestimated by as much as 104 percent. With a larger item calibration sample size (N = 2000), the overestimation was not nearly as great, although the overestimation approached 50% for middle ability levels.

2. With the smaller ratio of bank size to test length of 4:1, overestimation of the test information function occurred but the magnitude of the overestimation was reduced. With a low item calibration sample size (N = 500), overestimation was as

much as 63 percent. Whereas, with the larger item calibration sample size (N = 2000), overestimation did not exceed 26 percent.

Ratio of Item Bank Size to Test Length. Figure 3 shows the ratio of item bank size to test length to be an influential factor affecting the magnitude of overestimation of test information functions. Though, as was the finding with the optimal item selection procedure, for the situations simulated, the impact was less than sample size. The two main findings with respect to the ratio of bank size to test length were:

1. When the ratio of bank size to test length was 4:1 and using a lower sample size (N = 500), the size of overestimation was as much as 63 percent. However, the magnitude of overestimation was as great as 104 percent when the ratio was increased to 8:1.

2. Similar trends were present with the larger calibration sample (N = 2000). Again, the magnitude of the difference was about two thirds larger when the ratio was 8:1 compared to 4:1.

A comparison of the relative efficiency functions in Figures 2 and 4 shows the magnitude of overestimation to be greater in those tests constructed using content optimal item selection procedures compared to the optimal item selection procedures.

Conclusions

The results of our investigation document clearly the impact of the size of the examinee sample used in item calibration and the ratio of item bank size to test length on the accuracy of test information functions. At least for the conditions studied, examinee sample size was the more important of the two factors but, clearly, both factors were important. Again, at least for the situations simulated, the imposition of content specifications resulted in major inaccuracies in the test information

functions. Specific findings in this study are probably of limited interest - they are unique to the simulations. They are also somewhat unstable, especially for low and high ability levels where the generally short tests produced low levels of information. But, there is a main general finding from the study - test information functions tend to be overestimated because of "capitalizing on chance" in item selection and the amount of overestimation is due to factors investigated in this study.

There are several implications for practice: (1) tests do <u>not</u> perform as well as expected when the "best" items are selected to match a target test information function, and (2) standard errors are, correspondingly, under-estimated (assuming that the first set of values is taken as "true values") and so over-confidence in ability scores will result. If, on the other hand, the test developer's intention is to recalibrate the item statistics based on an actual test administration (which is common), one result will be that the updated test information function will be lower than was expected or desired. These results provide rather dramatic evidence of the influence of selecting the "best" items from an item bank to make up a test. Also, the size of the effect depends both on the number of examinees used in item calibration, and the ratio of the number of items in the bank relative to the length of the desired test.

At least two steps can be taken to reduce the problem:

1. Use large samples in item calibration to gain precision in item parameter estimates. An increase in the precision of item parameter estimates will reduce the significance of "capitalizing on chance."
2. Depending on the sample size used in item parameter estimation, exceed the desired target information function by (at least) 10% to 20%.

If one or both of the above suggestions are implemented, the problem associated with using over-estimated item parameter values in ability and standard error estimation can be reduced.

Table 2

Descriptive Statistics for the Item Parameter Estimates in Each Optimal Test

Bank Size/ Test Length Ratio	Examinee Sample Size	Test Length	Sample	Item Parameter b $\bar{X}$	b SD	a $\bar{X}$	a SD	c $\bar{X}$	c SD
4:1	500	20	A	.12	.74	1.12	.18	.17	.06
4:1	500	20	B	.13	.85	1.01	.26	.18	.06
4:1	2000	20	A	.28	.64	1.07	.17	.17	.08
4:1	2000	20	B	.27	.66	1.04	.20	.16	.08
8:1	500	10	A	-.15	.85	1.19	.17	.18	.07
8:1	500	10	B	-.14	.94	1.12	.25	.19	.06
8:1	2000	10	A	.15	.62	1.10	.10	.16	.07
8:1	2000	10	B	.19	.62	1.05	.19	.18	.08

Table 3

Descriptive Statistics for the Item Parameter Estimates in Each Content Optimal Test (Content Category 1 - 20%, 2 - 40%, 3 - 20%, 4 - 20%)

Bank Size/ Test Length Ratio	Examinee Sample Size	Test Length	Sample	Item Parameter b X	b SD	a X	a SD	c X	c SD
4:1	500	20	A	-.07	.79	1.04	.23	.17	.06
4:1	500	20	B	-.24	.79	.82	.26	.20	.05
4:1	2000	20	A	.05	.79	.94	.17	.16	.08
4:1	2000	20	B	-.11	.90	.83	.21	.16	.07
8:1	500	10	A	.16	.85	1.11	.20	.14	.05
8:1	500	10	B	-.26	.95	.84	.31	.19	.04
8:1	2000	10	A	.00	.87	.99	.18	.14	.07
8:1	2000	10	B	-.32	.97	.83	.27	.14	.06

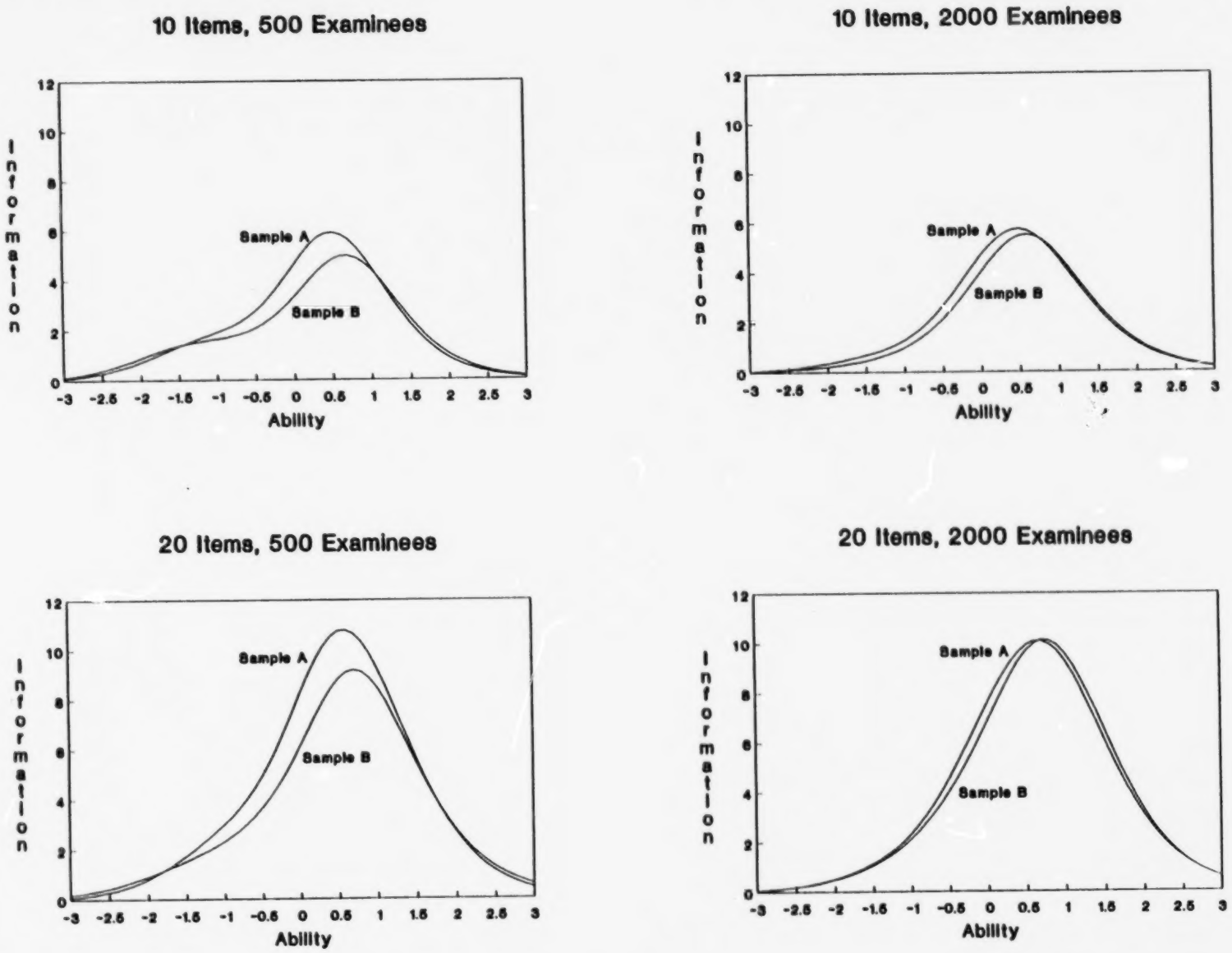
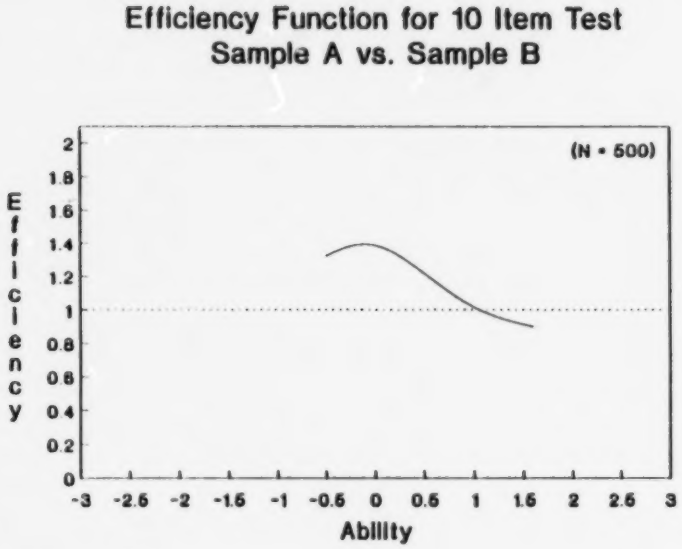
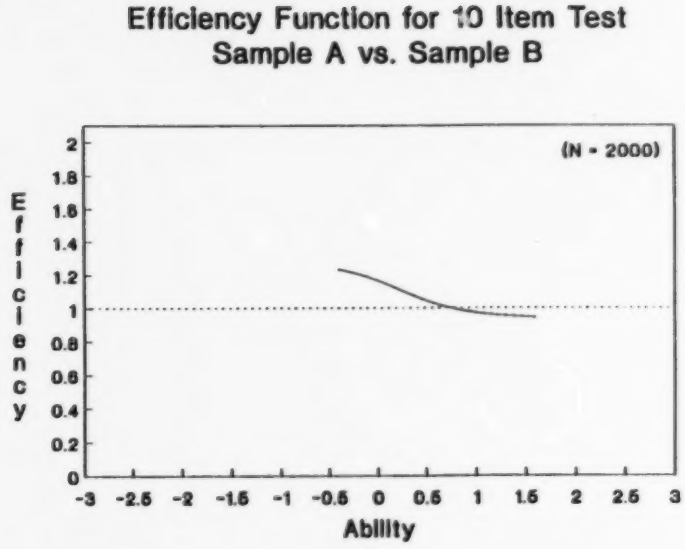
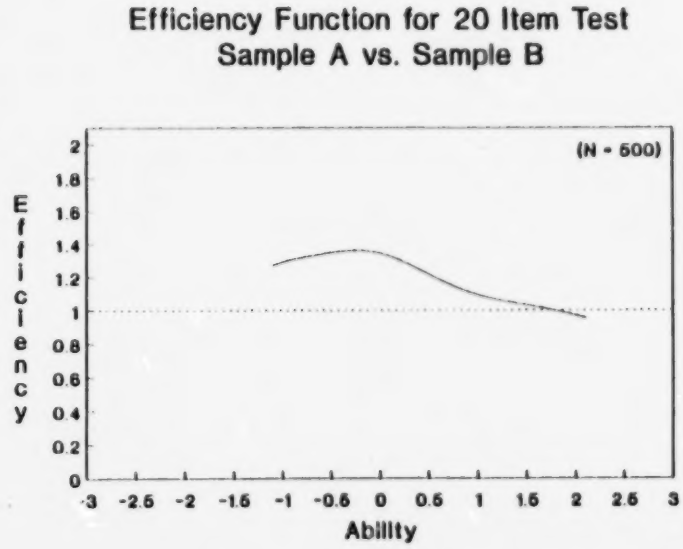


Figure 1. Test Information Functions for Tests Constructed Using Optimal Item Selection







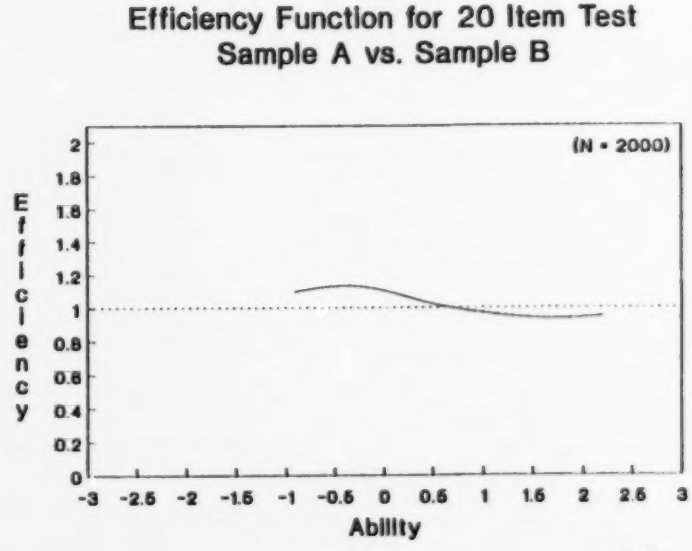


Figure 2. Relative Efficiency Functions for Tests Constructed Using Optimal Item Selection

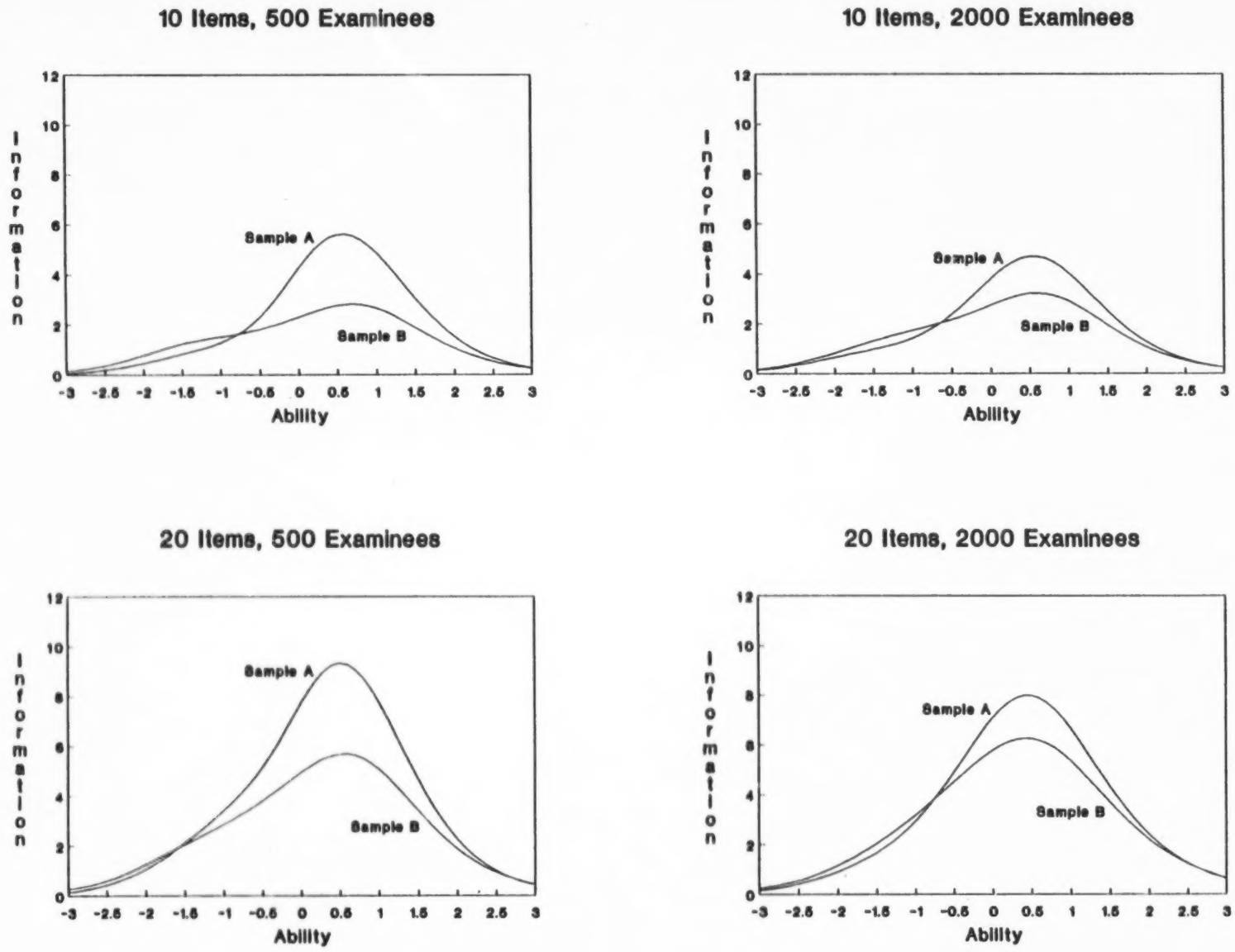


Figure 3. Test Information Functions for Tests Constructed Using Content Optimal Item Selection

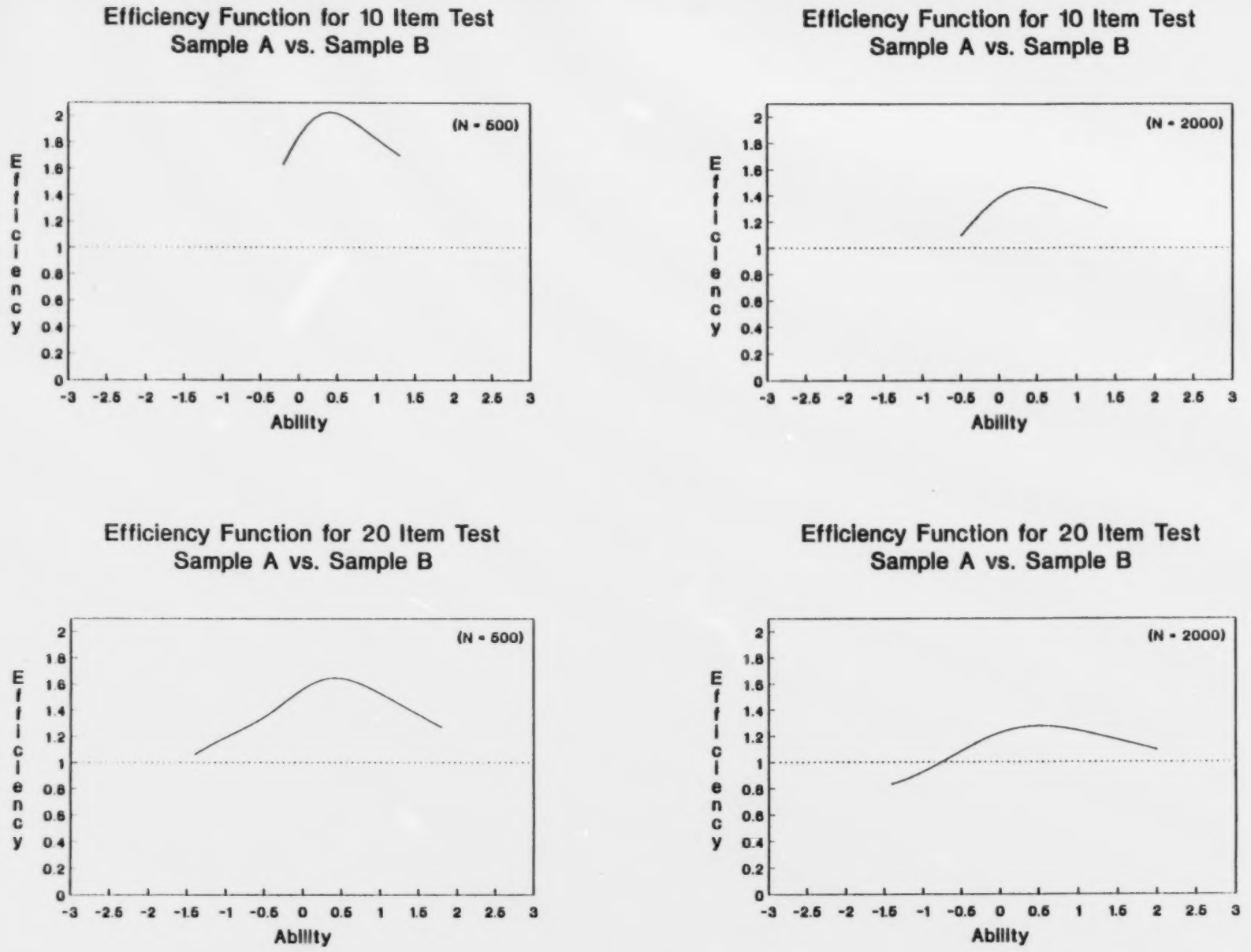


Figure 4. Relative Efficiency Functions for Tests Constructed Using Content Optimal Item Selection

References

Bezruczko, N., & Reynolds, A. J. (1987). Minimum proficiency skills tests: 1987 item pilot report (Citywide Report No. 87-1). Chicago: Chicago Public Schools.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), Educational measurement (3rd edition, pp. 147-200). New York: Macmillan.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (in press). Influence of item parameter estimation errors in test development. Journal of Educational Measurement.

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte-Carlo study. Applied Psychological Measurement, 6, 249-260.

Kingston, N., Leary, L., & Wightman, L. (1988). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admission Council.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Bock, R. D. (1991). BILOG 3: Item analysis and test scoring with binary logistic test models (2nd ed.). Mooresville, IN: Scientific Software, 17, 73-74.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T. J. J. M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-389.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. Psychometrika, 55(2), 371-390.

van der Linden, W. J. T., & Boekkooi-Timminga, W. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.

Verschoor, A., & Theunissen, T. J. J. M. (1991). Optimal test design (a software package). Arnhem, The Netherlands: CITO.

Table 1

Descriptive Statistics of the Item Parameter Estimates
in the Four Item Banks

Examinee Sample Size	Sample	Item Parameter b $\bar{X}$	b SD	a $\bar{X}$	a SD	c $\bar{X}$	c SD
2000	A	0.104	1.157	0.805	0.283	0.203	0.078
	B	0.133	1.110	0.817	0.319	0.199	0.082
500	A	0.178	1.098	0.840	0.267	0.199	0.063
	B	0.103	1.077	0.846	0.289	0.208	0.063